Journal of Hydroinformatics | 05.1 | 2003

Neural network rainfall-runoff forecasting based on continuous resampling

Robert J. Abrahart

ABSTRACT

Most neural network hydrological modelling has used split-sample validation to ensure good out-of-sample generalisation and thus safeguard each potential solution against the danger of overfitting. However, given that each sub-set is required to provide a comprehensive and sufficient representation of both environmental inputs and hydrological processes, then to partition the data could create limited individual representations that are, in some manner or other, deficient with respect to fitness-for-purpose. To address this issue a comparison has been undertaken between neural network rainfall-runoff models developed using (a) conventional stopping conditions and (b) a continuous single-model bootstrap. The results exhibit marginal improvement in terms of greater accuracies and better global generalisations—but the operation itself demonstrates substantial benefits through the provision of additional diagnostic capabilities and increased automation with respect to certain problematic aspects of the model development process.

Key words | bootstrap, neural network, rainfall-runoff forecasting

Robert J. Abrahart
School of Geography,
University of Nottingham,
Nottingham NG7 2RD,
UK
Tel: +44 115 84 66145;
Fax: +44 115 95 15249;
E-mail: *bob.abrahart@nottingham.ac.uk*

INTRODUCTION

The 1990s witnessed the advent and successful application of several innovative technologies in the field of hydrological modelling. This included: (i) the use of smart or soft computing methodologies; and (ii) the introduction of computer-based tools that made little or no explicit use of traditional mathematical symbols (Abbott 1999; Minns 2000). The investigation of neural solutions was a popular research endeavour and some reflections on their initial uptake can be found in compendium works such as: (i) Maier & Dandy (1999), (ii) ASCE (2000*a*,*b*) or (iii) Dawson & Wilby (2001). Streamflow prediction and forecasting received the most attention, since this problem is well suited to a neural solution, given the non-linear nature of the rainfall-runoff relationship and ease of access to long historical series of both precipitation and discharge data. For a comprehensive discussion on neural network terms and issues the reader is directed to selected texts such as Bishop (1995), Haykin (1999) or Reed & Marks (1999).

Most neural network hydrological modelling has adopted split-sample validation to ensure good out-of-sample generalisation and thus safeguard each solution against the danger of overfitting. However, given that each sub-set is required to provide a comprehensive representation of both environmental inputs and hydrological processes, then to partition the data could create limited individual representations that are, in some form or other, deficient with respect to fitness-for-purpose. This problem of reduced information content will be applicable to both model-construction and model-validation data sets and different selection options and combination strategies could lead to alternative modelling outcomes. The requirement for sub-division will also be a critical factor for small data sets and in situations where marked seasonal or annual variation exists.

To address this issue a comparison exercise was undertaken between neural rainfall-runoff models developed using (a) conventional split-sample procedures and

(b) continuous single-model bootstrapping. In each case a test data set was retained for 'proof of concept' evaluation purposes although the ultimate objective was to develop an efficient method that overcomes the traditional requirement for data splitting. These neural solutions were designed to forecast discharge on the Upper River Wye in Central Wales. Each neural bootstrapping operation was based on a continuous process of data selection and parameter adjustment, using small random sub-samples wherein each sub-sample was a random sample taken with replacement from the available hydrological record, in direct contrast to the standard method of model development based on large static sub-sets.

EXPERIMENTAL DESIGN

Problem of division

The recommended procedure for evaluating the performance of a neural model is to split the available data into: (i) a training set that is used for parameter estimation based on gradient descent against some cost function; (ii) a validation set that is used to monitor performance, to determine a stopping point after which the solution becomes overfitted, or to set additional parameters or hyper-parameters such as weighted penalties on over-complex models; and (iii) one or more test sets. The data sets in a split-sample approach share no patterns in common and each set is expected to provide an adequate representation of the problem space in terms of range and completeness. Each set must also encapsulate the relevant characteristics and covariance of each input distribution and output distribution, together with the assemblage of complex interwoven deterministic relationships, that exists between them.

There is no authoritative method that can be used to divide the data, or to confirm that each split sample is a good representation, and several different approaches have been adopted in the past, e.g. random samples, use of standard temporal units such as annual data sets, or division based on equivalent statistical descriptors such as measures of centralisation and dispersion. The best word of advice on split-sample modelling is to use large samples, in the expectation that sufficient information will be contained within each data set, since larger data sets will often provide more accurate approximations (Reed & Marks 1999). For an illustrative discussion on the potential pitfalls of ignoring variation across static divisions, or the danger of drawing strong conclusions from modelling with static divisions that exhibit marked sensitivities to data splitting, see LeBaron & Weigend (1998).

Bootstrap manoeuvre

The bootstrap (Efron 1979; Efron & Tibishirani 1993) is a computational procedure that uses intensive resampling, with replacement, to reduce uncertainties. The aim of resampling is to mimic the random component of a process and to reduce variance through averaging over numerous different partitions of the data. However, the decision on which item(s) is (are) to be resampled is a multifaceted issue that must be determined from a consideration of the stochastic component of the modelling process (Moony & Duval 1993), e.g. components, coefficients or residuals. The bootstrap mechanism is often used to process hundreds or thousands of subsets, such that an empirical estimate of a specified output distribution is produced, and from which certain fundamental characteristics of the population can be calculated, e.g. means, variances or cumulants. It can also be used to produce statements about probabilities, to generate inferences about true parameters, or to determine confidence intervals.

The use of non-parametric bootstrap approaches in hydrological modelling is on the increase. Documented applications range from estimating means, confidence intervals, or parameter uncertainties to network design techniques (e.g. Cover & Unny 1986; Tasker 1987, 1999; Woo 1989; Moss & Tasker 1991; Zucchini & Adamson 1989; Di Stefano *et al.* 2000) and the adoption of more complicated block-based methodologies that endeavour to maintain temporal dependence or spatial covariance (e.g. Lall & Sharma 1996; Vogel & Shallcross 1996; Sharma *et al*. 1997; Tasker & Dunne 1997; Srinivas & Srinivasan 2000, 2001). The application of bootstrap methodologies to build neural solutions is also the subject

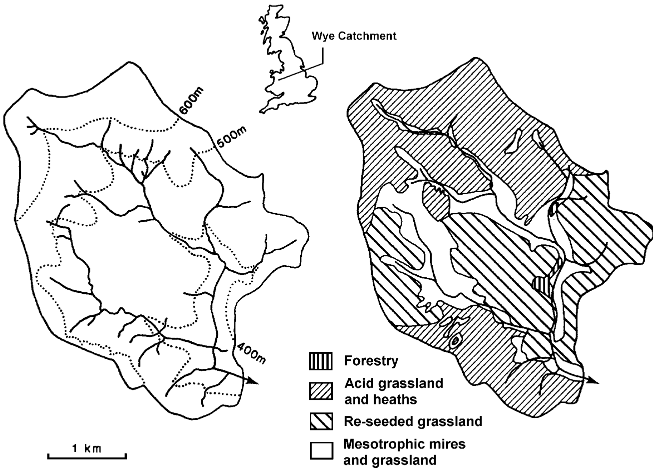

Figure 1 | Upper River Wye catchment (after Beven *et al.* 1984).

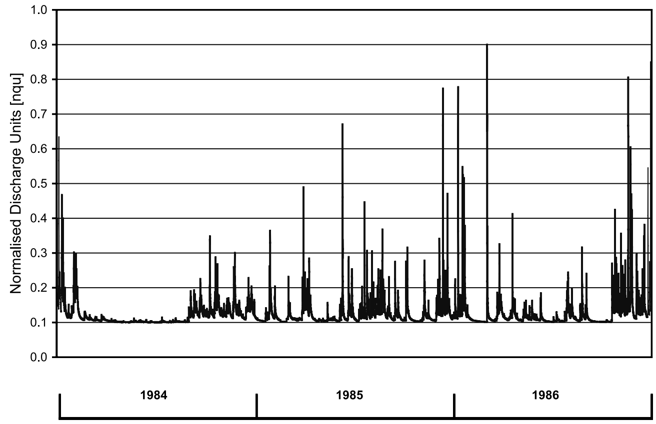

Figure 2 | Hydrograph for the Upper River Wye 1984–86.

of current research. There are two natural paths for randomness to enter a neural model-building operation: through different choices about splitting the data, or through different choices about network initialisation, architecture and training. Either path, or both paths together, can be bootstrapped. The neural bootstrap has been used to perform bootstrap aggregation (bagging) of multi-model ensembles which produced averaged outputs and a more stable solution (Hsieh & Tang 1998; Tang *et al*. 1998) and bootstrap assessment of multi-model multi-data solutions which established the influence of different components (LeBaron & Weigend 1998). More sophisticated neural bootstraps have also been used to estimate confidence bounds for network outputs (Efron & Tibshirani 1993) and for bootstrapping residuals (i) to evaluate forecasting power (Weigend *et al*. 1992) and (ii) to obtain error bars on iterated time series predictions (Connor 1993).

Hydrological data

The Upper River Wye basin in Central Wales was selected for these investigations (Figure 1). This is a small upland research catchment that has moderate spatial variation and a quick response. The basin covers an area of 10.55 km^2, elevations range from 350–700 m and average annual rainfall is 2500 mm. Previous hydrological modelling of this catchment includes Beven *et al.* (1984), Bathurst (1986), Quinn & Beven (1993), Abrahart & Kneale (1997) and Abrahart *et al.* (1999). Discharge (Q) and rainfall (R) data were available on a one-hour time step for the period 1984–86. Figure 2 depicts variation in discharge: 1984 had a summer drought; 1985 contained a good spread of events; 1986 showed greater divergence and experienced the biggest floods.

Modelling predictors were identified using the 'pick-and-mix' significant relationships approach of Dawson & Wilby (1998). To obtain maximum forecasting power, from a minimum set of inputs, correlation analysis was performed against lags and moving averages of rainfall and discharge to ascertain which factors would be the strongest predictors of current discharge (Q). This use of lags and moving averages provided short-term recollection of previous events and antecedent conditions. Further, using objective tools to search for suitable 'input drivers' is equivalent to the identification of catchment parameters and such operations must be distanced from issues associated with the division of data between model building and model testing operations. In a similar fashion this approach also parallels the process of model selection for a particular location or problem. Thus correlation was performed on the full data set although purists might argue that the test set should have been excluded from such operations. There was, in addition, a practical reason for doing this: it was difficult to obtain agreement between the various possible combinations of paired annual series on

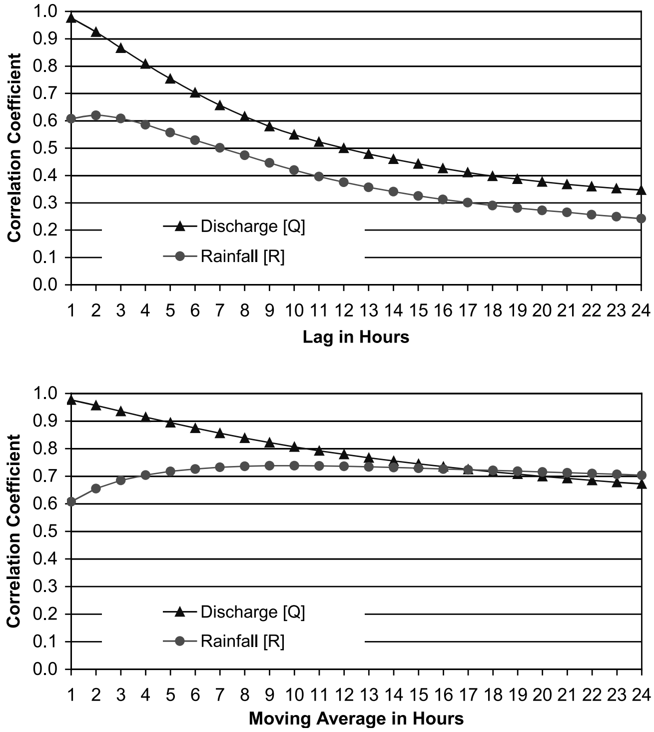

Figure 3 | Correlation analysis of lags and moving averages against predictand.

Table 1 | Correlation matrix of selected predictors against predictand

Modelling output	Modelling input				
	Q_{t-1}	R_{t-2}	$R_{avg[10]}$	sin (CLOCK)	cos (CLOCK)
Q_t	0.9764	0.6205	0.7384	− 0.0980	0.2478

what did or did not constitute the best set of short-term 'input drivers'.

Figure 3 contains plots of the correlation coefficients from which the optimal inputs were identified as Q_{t-1}, R_{t-2} and $R_{avg[10]}$. The plot of moving average discharge exhibited a progressive degradation and was omitted from further consideration since the highest value, $Q_{avg[1]}$, is equivalent to Q_{t-1}. Two additional 'drivers' were added to prevent excessive generalisation and to allow for non-linearities in modelling response: sin (CLOCK) and cos (CLOCK). These inputs, derived from annual hour count (CLOCK), can discover and incorporate seasonal or annual influences (Abrahart *et al.* 2001)—which is important since an agricultural catchment might be expected to produce different responses in summer (drier) and winter (wetter). Table 1 provides correlation statistics between each individual predictor and the predictand. To overcome problems associated with upper-limit and lower-limit saturation the input and output data were standardised, using a linear transformation, to an intermediate range (0.1–0.9). For simplification purposes all results will be reported in normalised discharge units (nqu).

Standard approach

Two standard solutions were developed using a 5:5:1 backpropagation network with sigmoid transfer functions and random initialisation (between plus and minus one). Selection of an optimal architecture is problematic but previous neural network rainfall-runoff research has demonstrated that: (i) acceptable models can be produced from standard solutions of modest size; (ii) a large number of hidden units has little or no real impact on the end result; and (iii) the benefit of multiple hidden layers is marginal in comparison to the numerical overheads involved (Minns & Hall 1996; Abrahart & See 2000). These findings correspond to empirical investigation into the effectiveness of different methods of ensemble creation (an 'ensemble' is a combination of redundant networks) which suggests that variation in the training data has the greatest potential for creating networks that produce different errors (Sharkey & Sharkey 1995; Sharkey *et al.* 1996; Tumer & Ghosh 1996). It is also commensurate with the opinion that neural networks will, in most cases, attempt to build an identical function from a given set of data, albeit that alternative degrees of generalisation, or different levels of sub-optimal solution, are possible (Sharkey 1999).

The processed data were split into annual data sets and two standard runs were undertaken to provide a comparison against which the bootstrap results could be evaluated. The role and function of each annual data set during each model development operation was as follows:

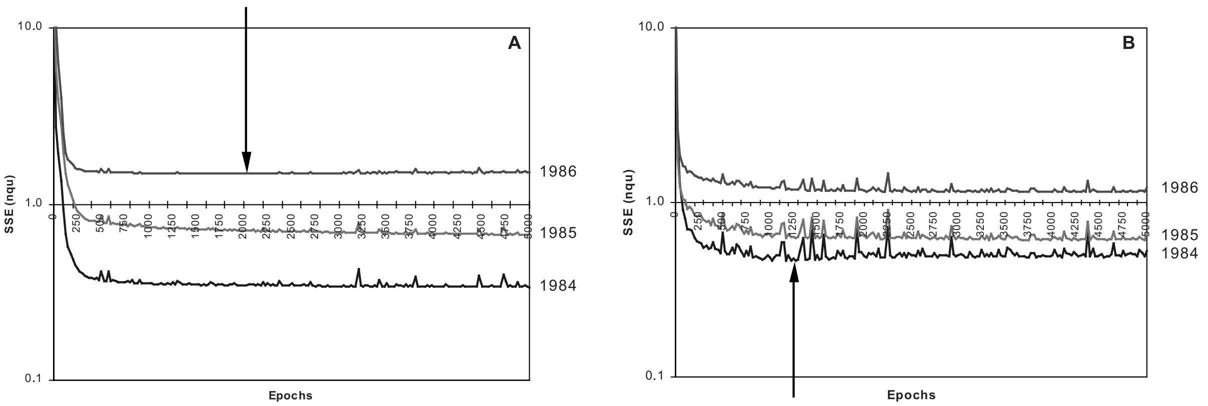

Figure 4 | Model selection based on split-sample validation for RUN-A and RUN-B: log scale used to obtain maximum differentiation; arrows indicate point of inflection and optimal solution.

RUN-A: 1984 training set; 1986 validation set; 1985 test set

RUN-B: 1986 training set; 1984 validation set; 1985 test set

These two organisational groupings were based on visual inspection of the hydrographic record. To encapsulate a full range of outliers and conditions, the construction process needed to include the summer drought and the largest floods, so solutions were developed on paired combinations of the annual data sets for 1984 and 1986. Further, using role reversal, these temporal divisions can be used in different modes to build alternative modelling solutions. In operation (A) 1986 data provided a stopping condition to prevent overfitting on the 1984 model; and in operation (B) their implementation was reversed. The central period, 1985, comprised intermediate catchment conditions and contained a large number of flood events. It was, in consequence, a good test set that had no requirement for questionable extrapolation of the predictand.

Low rates of learning (0.2) and momentum (0.1) were applied throughout. The training data were presented in random order and sum squared error statistics computed at regular intervals on each annual data set. These results were then translated into a combined graph from which the optimal modelling solution, in each experiment, could be determined. Models were selected at the point of inflection on the validation error curve; error associated with the validation data set was thereafter observed to increase, in a progressive manner, which is indicative of overfitting. The optimal solutions were obtained at (A) 2,000 epochs and (B) 1,275 epochs (Figure 4).

Bootstrap simulation

Most neural bootstrap operations have, to date, involved building a large number of networks–one for each set of resampled data. Each model is developed in a standard manner and the output related to each set of inputs at each instant collated, such that means or standard deviations can then be computed and used to describe the output distribution of either predictions or errors. This process is said to produce a stable mean, which is not subject to the vagaries of split-sample validation, and offers a measure of reliance in terms of potential variation. However, descriptors of centralisation and dispersion provide a scale of correspondence, but it is not a true 'confidence region' in terms of predicted modelling output. For a method to estimate true confidence regions in the form of local error bars that depend upon relative location in input space see Nix & Weigend (1995). The adoption of an ensemble solution is also problematic, since this involves extensive duplication of the model building process, with no clear separation between random data selection and random model development.

The computational effort that must be expended to train and test thousands of solutions in an automated manner is a realistic option. But to maintain the tradition of split-sample validation is to risk an accumulation of the methodological drawbacks and practical problems that are associated with 'stopped training'. The main criticisms of 'stopped training' are listed in Sarle (2001): rules of thumb on (i) the number of cases in each set of data; (ii) the split of data into training and validation sets using either random selection or some form of systematic algorithm; or (iii) the decision on when validation error 'starts to increase' since it could go up or down numerous times during training. The safest method is to train to convergence, then go back and determine which iteration had the lowest validation error (as used in the standard approach). For more elaborate algorithms see Prechelt (1994, 1998). Last, but not least, neither data set makes full use of the entire sample and standard statistical theories or constructs are not applicable in this practical working *modus operandi*.

To examine alternative approaches a single-model-bootstrap solution has been designed and implemented. This solution is based on resampling with replacement in which the model is built from a continuous sequence of resampled data. The model comprises four individual programs that are organised as loose-coupled components: (i) master control program; (ii) data resampling program; (iii) output interrogation program; and (iv) a neural network simulator. The pseudo-random number generator that was used in the bootstrap resampling procedure was RAN-2 (Press *et al*. 1993). The modelling operation worked as follows:

1. Extract small random sub-sample from the main model building data set.
2. Train network with sub-sample and perform a limited amount of weight adjustment.
3. If desired then:
   - test the solution
   - update mean and standard deviation outputs for each forecast in the test data set.
4. Do until told to stop:
   - repeat steps 1–4

Each random sub-sample will attempt to produce a solution that is, in some manner or other, unique to itself and to the information that is contained within its data pairs. Each set of extracted data within this modelling operation is, for that reason, in direct competition with all other sets of extracted data, such that the forecasting solution which is being developed will attempt to address a series of different individual biases, one for each set of data that is used to train it. The upshot of this 'battle to capture the solution' is that each random sub-sample will influence the level of generalisation, through the process of construction, in which small changes are applied to the modelling solution over time in the sprit of competition and progressive smithing. Thus emerges a 'shifting average' that has no undue allegiance to the specifics of a single annual series, but which approximates the fundamental properties or common responses of the resampled data, albeit inclined towards a stronger representation of the most recent random selection. Moreover, since the model fitting operation is continuous, there is no end to this process and no single final product, which means that the dilemma of 'stopped training' is avoided as opposed to being proliferated or compounded. Earlier neural bootstrap studies have also revealed that variation in forecasts due to changes in structure or architecture are small in comparison to those that arise from sample splitting (LeBaron & Weigend 1998). Thus neither architecture nor modelling parameters were bootstrapped.

To maintain commonalities with the standard approach an identical 5:5:1 backpropagation architecture was used and random resampling was applied to an amalgamated database that comprised all patterns in the annual data sets for 1984 and 1986. It could be argued that this use of two annual data series provides an unfair advantage; but the whole point of this modelling exercise is to avoid the traditional problems of squandered resources and natural bias towards one or other of the split-sample data sets. To help steer clear of unwarranted overfitting against a single set of resampled patterns the backpropagation parameters in this continuous sequence of data selection and model adjustment were decreased. The operational environment of minimum adjustment for each random sub-sample was set at: epochs = 100; learning = 0.1 and momentum = 0.05. There is no logical

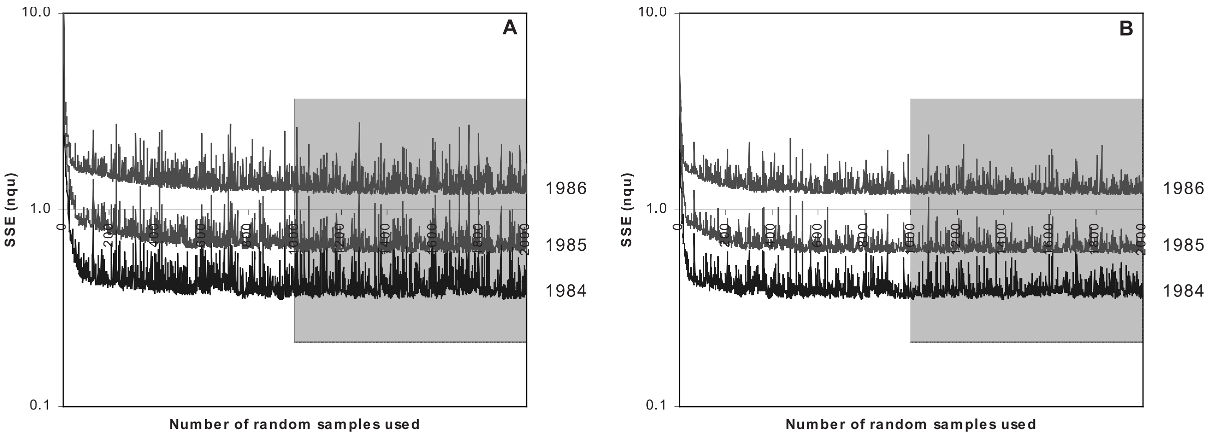

Figure 5 | Error plots for bootstrap modelling (A) BTSP-500 and (B) BTSP-1000: log scale used to obtain maximum differentiation; shaded area indicates period of extraction.

stopping point (or related problem) in a continuous simulation which switches from one focus to another at each step in the process of data selection and model adjustment, so the model outputs for each forecast in each annual series were extracted and summarised over time (not over optimised solutions) using means and standard deviations. It should be noted that aggregate measures, such as means and standard deviations, can be calculated on a continuous basis and are thus appropriate for continuous modelling operations, whereas the provision of alternative statistics from a comprehensive sequence of all past records, such as medians or percentile bandwidths, is not a practical computational proposition. Model development based on two different random sample sizes was investigated (BTSP-500, BTSP-1000) and in each case summation statistics for each forecast in the test set were computed over a fixed period of post-initial solution development resampling operations (1,000–2,000). Error plots for the two continuous modelling investigations are provided in Figure 5.

EMPIRICAL RESULTS

It is important to consider a number of statistical evaluators since there is no single definitive measure that can determine the success of each forecast (Houghton-Carr 1999; Legates & McCabe 1999; Hall 2001). Eight numerical descriptors were therefore computed:

- coefficient of efficiency[1] (COE)
- root mean squared error (RMSE)
- maximum under-prediction (MUP)
- maximum over-prediction (MOP)
- largest positive change in discharge (LPC)
- largest negative change in discharge (LNC)
- numerical range of change in discharge (ROC)
- standard deviation of change in discharge (SDC)

Diagnostic outputs for selected data sets are provided in Tables 2, 3 and 4. Forecasts were also subject to visual inspection and graphical analysis, to provide qualitative

[1]Nash & Sutcliffe (1970).

Table 2 | Numerical results for standard solutions (test set in bold)

Solution	Data set	COE	RMSE	MUP	MOP
RUN-A	1984	0.9668	0.0063	−0.2646	0.0829
	1985	**0.9366**	**0.0090**	**−0.2037**	**0.1642**
	1986	0.9462	0.0130	−0.2085	0.1951
RUN-B	1984	0.9552	0.0073	−0.2455	0.1225
	1985	**0.9440**	**0.0085**	**−0.2018**	**0.1676**
	1986	0.9574	0.0116	−0.2019	0.2001

Table 3 | Numerical results for bootstrap solutions

Solution	Data set	COE	RMSE	MUP	MOP
BTSP-500	1985	0.9452	0.0084	– 0.2049	0.1567
BTSP-1000	1985	0.9456	0.0083	– 0.2041	0.1556

Table 4 | Numerical results for predicted change in discharge

Solution	Data set	LPC	LNC	ROC	SDC
ORIGINAL	1985	0.2066	– 0.1658	0.3723	0.0086
RUN-A	1985	0.0537	– 0.0978	0.1515	0.0031
RUN-B	1985	0.0769	– 0.0444	0.1213	0.0035
BTSP-500	1985	0.0383	– 0.0404	0.0788	0.0022
BTSP-1000	1985	0.0336	– 0.0380	0.0715	0.0019

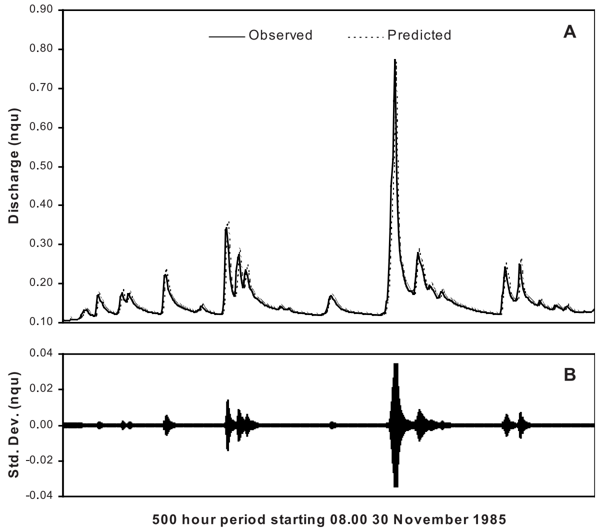

Figure 6 | BTSP-1000 time series plot for biggest winter event in test data set: (A) observed and predicted discharge, (B) standard deviation of bootstrap forecasts.

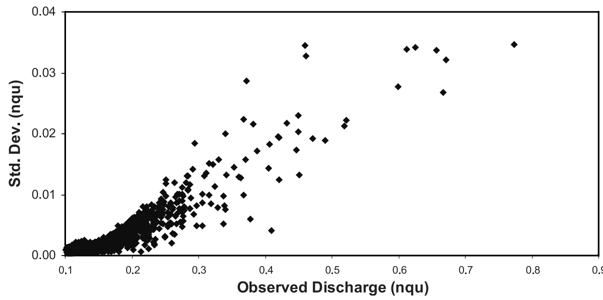

Figure 7 | BTSP-1000 scatterplot to illustrate relationship between standard deviation of bootstrap forecasts and observed discharge for test data set.

information about temporal performance and error characteristics, and to facilitate a detailed examination of the relationship between modelling difficulties and bootstrap standard deviation indices. Pertinent graphics are provided in Figures 6 and 7.

DISCUSSION

The bootstrap manoeuvre can be used to counteract numerous difficulties that arise from the haphazard process of model development, through the construction of ensemble solutions, and related multi-model output averaging. However, in the reported research, a continuous single-model bootstrap has been developed to exploit the untapped benefits of progressive construction, which uses on-going competition between resampled sub-sets, to establish an automated mechanism that will produce an optimal solution averaged over time.

The four neural solutions produced an excellent set of statistical results and in all cases BTSP-1000 did a little bit better than BTSP-500. Further, there are no signs of potential overfitting, and no indication of problematic sub-optimal traps. COE and RMSE statistics indicated better levels of global generalisation for the bootstrap operations, although in real terms the difference between the four neural solutions was slight, and the bootstrap models were not much better than the superior member of their two standard counterparts. The greatest errors occurred under similar hydrological circumstances and such problems are thought to have arisen from a combination of factors that include: (i) deficiencies in the modelling record and (ii) failure to provide an adequate mechanism for the extrapolation of complex hydrological relationships.

MUP and MOP, are associated with problematic situations in which the forecasts appeared to be an hour or so out-of-step, in terms of lateness on both rising limbs (under-prediction) and falling limbs (over-prediction) for individual events–termed 'phase error'. This result is perhaps to be expected since there will be a limited degree of explanation associated with the selection of a minimal set of parsimonious inputs, since there will be under-sampling of rising limbs in comparison to recession curves which form a larger proportion of the temporal record, and because there were no 'drivers' which allowed for the cessation of rainfall to be anticipated. The bootstrap solutions, in comparison to their split-sample counterparts, were observed to provide a similar (albeit poorer) response on steeper sections of the rising limb and a much better response on shallower sections of the falling limb. So, although the bootstrap operation has in overall terms created a more generalised solution, under certain circumstances a more generalised solution will produce a weaker response, which leads to similar or greater errors.

Table 4 reveals substantial problems in the degree of change, between current and predicted discharge, at each time step. This feature is not well modelled and the bootstrap solutions, which provided the greatest level of generalisation, exhibit the lowest range and weakest potential reaction. It could be argued that the neural solutions are taking too much notice of current discharge such that extra 'drivers' are needed to build a better model. However, it is also possible to contend that sharp increases and decreases appear as isolated energetic events, such that forecasting minor changes from one moment to the next is, in fact, the correct response for a mechanism that endeavours to offer a robust universal approximation. To investigate this issue, the research agenda must therefore be extended to a consideration of biased distributions and histogram equalisation operations based on change in discharge (global skewness 4.66; global kurtosis 158.90), perhaps in some manner or other related to the class of procedures that are used to transform skewed data into a normal distribution.

Figure 6 provides an illustration of these problems. It also demonstrates the nature of the relationship between the bootstrap standard deviation indices and the hydrographic record at each point. Higher predictions and major changes are observed to be associated with greater variation in forecasting output. Further, strong numerical relationships were identified, based on correlation analysis of the standard deviation indices against discharge (BTSP-500, $R = 0.8858$; BTSP-1000, $R = 0.8767$) and absolute change in discharge (BTSP-500, $R = 0.6887$; BTSP-1000, $R = 0.6886$). These collective observations suggest that bootstrap modelling was biased towards the lower levels of discharge and weaker changes; there was a profusion of mediocre samples or patterns such that the most significant hydrological features were treated as outliers and not the norm. To investigate this issue the research agenda must also be extended to a consideration of biased distributions and histogram equalisation operations based on discharge (global skewness 6.82; global kurtosis 74.48).

CONCLUSIONS

- The concept of continuous model building has been demonstrated. This work has substantial implications for real-time forecasting and the need to develop solutions that adapt to changing circumstances, based on fresh input data, as and when the need arises.
- The reported bootstrap mechanism provided marginal improvements in terms of greater accuracies and better global generalisations–but with diminished response in more challenging situations. Such failings can be attributed to representational inequalities in the original data and illustrates a need for efficacious transformation tools or procedures.
- The main benefits were observed to be associated with increased automation in respect of (i) a reduction in guesstimates on the division of data and (ii) a release from the need to select an optimum modelling solution–based on one or more user defined parameters or conditions.
- The standard deviation and degree of change indices provided useful diagnostic tools. Explicit information was obtained about difficulties on: (i) higher magnitude prediction; (ii) representation of

rapid change; and (iii) potential swamping from an unwarranted number of mundane patterns.

- Innovative software solutions must be developed (i) to perform multifaceted histogram equalisation and (ii) to provide alternative inputs that produce higher temporal accuracies.
- Further research is needed to investigate block bootstrapping and confidence intervals.

ACKNOWLEDGEMENTS

Stuttgart Neural Network Simulator (SNNS Group (1990–98)). Upper Rive Wye discharge and rainfall data (Centre for Ecology and Hydrology, Wallingford, UK).

REFERENCES

Abbott, M. B. 1999 Introducing hydroinformatics. *J. Hydroinformatics* **1**(1), 3–19.

Abrahart, R. J. & Kneale, P. E. 1997 Exploring neural network rainfall-runoff modelling. *Proceedings 6th National Hydrology Symposium, University of Salford, 15–18 September 1997*, 9.35–9.44.

Abrahart, R. J. & See, L. 2000 Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrol. Processes* **14**, 2157–2172.

Abrahart, R. J., See, L. & Kneale, P. E. 1999 Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. *J. Hydroinformatics* **1**(2), 103–114.

Abrahart, R. J., See, L. & Kneale, P. E. 2001 Investigating the role of saliency analysis with a neural network rainfall-runoff model. *Comput. Geosci.* **27**(8), 921–928.

ASCE (ASCE Task Committee on the Application of Artificial Neural Networks in Hydrology) 2000*a* Artificial neural networks in hydrology. I: Preliminary concepts. *J. Hydrol. Engng.* **5**(2), 115–123.

ASCE (ASCE Task Committee on the Application of Artificial Neural Networks in Hydrology) 2000*b* Artificial neural networks in hydrology. II: Hydrologic applications. *J. Hydrol. Engng.* **5**(2), 124–137.

Bathurst, J. 1986 Sensitivity analysis of the Systeme Hydrologique Europeen for an upland catchment. *J. Hydrol.* **87**, 103–123.

Beven, K., Kirkby, M. J., Schofield, N. & Tagg, A. F. 1984 Testing a physically-based flood forecasting model (TOPMODEL) for three U.K. catchments. *J. Hydrol.* **69**, 119–143.

Bishop, C. M. 1995 *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Connor, J. T. 1993 Bootstrap methods in neural network time series prediction. In *International Workshop on Applications of Neural Networks to Telecommunications* (Eds J. Alspector, R. Goodman & T. X. Brown), pp. 125–131. Erlbaum, Hillsdale, NJ.

Cover, K. A. & Unny, T. E. 1986 Application of computer intensive statistics to parameter uncertainty in streamflow synthesis. *Wat. Res. Bull.* **22**(3), 495–507.

Dawson, C. W. & Wilby, R. L. 1998 An artificial neural network approach to rainfall–runoff modelling. *Hydrol. Sci. J.* **43**, 47–66.

Dawson, C. W. & Wilby, R. L. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geog.* **25**, 80–108.

Di Stefano, C., Ferro, V. & Porto, P. 2000 Applying the bootstrap technique for studying soil redistribution by caesium-137 measurements at basin scale. *Hydrol. Sci. J.* **45**(2), 171–184.

Efron, B. 1979 Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. & Tibshirani, R. J. 1993 *An Introduction to the Bootstrap*. Chapman & Hall, London.

Hall, M. J. 2001 How well does your model fit the data. *J. Hydroinformatics* **3**(1), 49–55.

Haykin, S. 1999 *Neural Networks: A comprehensive Foundation* 2nd edn. Prentice Hall, Englewood Cliffs, NJ.

Houghton-Carr, H. A. 1999 Assessment criteria for simple conceptual daily rainfall-runoff models. *Hydrol. Sci. J.* **44**(2), 237–261.

Hsieh, W. W. and Tang, B. 1998 Applying neural network models to prediction and data analysis in meteorology and oceanography. *J. Am. Meteorol. Soc.* **79**(9), 1855–1870.

Lall, U. & Sharma, A. 1996 A nearest neighbor bootstrap for resampling hydrologic time series. *Wat. Res. Res.* **32**(3), 679–693.

LeBaron, B. & Weigend, A. S. 1998 A bootstrap evaluation of the effect of data splitting on financial time series. *IEEE Trans. Neural Networks* **9**, 213–220.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Wat. Res. Res.* **35**(1), 233–241.

Maier, H. R. & Dandy, G. C. 1999 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Software* **15**, 101–123.

Minns, A. W. 2000 Subsymbolic methods for data mining in hydraulic engineering. *J. Hydroinformatics* **2**(1), 3–14.

Minns, A. W. & Hall, M. J. 1996 Artificial neutral networks as rainfall-runoff models. *Hydrol. Sci. J.* **41**(3), 399–417.

Mooney, C. Z. & Duval, R. D. 1993 *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage Publications, Newbury Park, CA.

Moss, M. E. & Tasker, G. D. 1991 An intercomparison of hydrological network-design technologies. *Hydrol. Sci. J.* **36**(3), 209–221.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models, Part 1: a discussion of principles. *J Hydrol.* **10**(3), 282–290.

Nix, D. A. & Weigend, A. S. 1995 Learning local error bars for non-linear regression. In *Advances in Neural Information Processing Systems 7—NIPS'94*, (Eds G. Tesauro, F. S. Touretzky & T. K. Leen), pp. 488–496. MIT Press, Cambridge, MA.

Prechelt, L. 1994 PROBEN1—A set of neural network benchmark problems and benchmarking rules. Technical Report 21/94, Universitat Karlsruhe, Germany. URL = ftp://ftp.ira.uka.de/pub/papers/techreports/1994/1994-21.ps.gz.

Prechelt, L. 1998 Early stopping—But when? In *Neural Networks: Tricks of the Trade,* (Eds G. B. Orr & K.-R. Mueller), pp. 55–69. Springer, Berlin.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1993. *Numerical Recipes in C*, 2nd edn. Cambridge University Press, Cambridge.

Quinn, P. F. & Beven, K. J. 1993 Spatial and temporal predictions of soil moisture dynamics, runoff, variable source areas and evapotranspiration for Plynlimon, Mid-Wales. *Hydrol. Proc.* **7**, 425–448.

Reed, R. D. & Marks, R. J. 1999 *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press, Cambridge, MA.

Sarle, W. S. 2001 FAQ document for Usenet newsgroup comp.ai.neural-nets. ftp.sas.com/pub/neural/FAQ.html.

Sharkey, A. J. C. (ed.) 1999 *Combining Artificial Neural Networks: Ensemble and Modular Multi-Net Systems*. Springer, London.

Sharkey, A. J. C. & Sharkey, N. E. 1995 *How to improve the reliability of artificial neural networks*. *Technical Report CS-95-11*, Department of Computer Science, University of Sheffield.

Sharkey, A. J. C., Sharkey, N. E. & Chandroth, G. O. 1996 Neural nets and diversity. *Neural Comput. Appl.* **4**, 218–227.

Sharma, A., Tarboton, D. G. & Lall, U. 1997 Streamflow simulation: a nonparametric approach. *Wat. Res. Res.* **33**(3), 291–308.

SNNS Group. 1990–98 *Stuttgart Neural Network Simulator, User Manual—Version 4.1*. www-ra.informatik.uni-tuebingen.de/SNNS/.

Srinivas, V. V. & Srinivasan, K. 2000 Post-blackening approach for modelling dependent annual streamflows. *J. Hydrol.* **230**, 86–126.

Srinivas, V. V. & Srinivasan, K. 2001 Post-blackening approach for modelling periodic streamflows. *J. Hydrol.* **241**, 221–269.

Tang, B., Hsieh, W. & Tangang, F. T. 1998 Neural network model forecasts of the NINO3.4 sea surface temperature. *Expt. Long-Lead Forecast Bull.* **7**, 1. grads.iges.org/ellfb/Mar98/tan.html.

Tasker, G. D. 1987 Comparison of methods for estimating low flow characteristics of streams. *Wat. Res. Bull.* **23**, 1077–1083.

Tasker, G. D. 1999 Bootstrapping periodic ARMA model to forecast streamflow at multiple sites. *Comput. Sci. Statist.* **31**, 296–299.

Tasker, G. D. & Dunne, P. 1997 Bootstrap position analysis for forecasting low flow frequency. *J. Wat. Res. Plan. Mngt* **123**(6), 359–367.

Tumer, K. & Ghosh, J. 1996 Error correlation and error reduction in ensemble classifiers. *Connect. Sci.* **8**, 385–404.

Vogel, R. M. & Shallcross, A. L. 1996 The moving blocks bootstrap versus parametric time series models. *Wat. Res. Res.* **32**(6), 1875–1882.

Weigend, A. S., Huberman, B. A. & Rumelhart, D. E. 1992 Predicting sunspots and exchange rates with connectionist networks. In *Nonlinear Modeling and Forecasting* (Eds M. Casdagli & S. Eubank), pp. 395–432. Addison-Wesley, New York.

Woo, M. K. 1989 Confidence intervals of optimal risk-based hydraulic design parameters. *Can. Wat. Res. J.* **14**, 10–16.

Zucchini, W. & Adamson, P. T. 1989 Bootstrap confidence intervals for design storms from exceedance series. *Hydrol. Sci. J.* **34**, 41–48.